DOCUMENT RESUME

ED 230 564 TM 830 220

AUTHOR Roid, Gale H.; Wendler, Cathy L. W.
TITLE Item Bias Detection and Item Writing Technology.
PUB DATE Apr 83
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Difficulty Level; Intermediate Grades; *Item Analysis; Item Banks; Multiple Choice Tests; Science Tests; *Sex Bias; Test Bias; *Test Construction; Test Format; *Test Items
IDENTIFIERS Distractors (Tests); *Parallel Test Forms

ABSTRACT

The development of the emerging technology of item writing was motivated in part by the desire to reduce potential subjectivity and bias between different item writers who attempt to construct parallel achievement tests. The present study contrasts four test forms constructed by the combined efforts of six item writers using four methods of item writing: (1) informal and subjective, (2) objective-based, (3) sentence transformation with writer's choice of distractors, and (4) sentence transformation with algorithmic creation of distractors. Three item bias detection techniques were used to examine potential sex bias in each test form: (1) delta plots of transformed item difficulty in male and female groups, (2) Scheuneman's chi-square-correct, and (3) internal consistency estimates (alpha) for each test form in each group. The basic data of the study were from 218 4th- and 5th-grade students who were randomly assigned the four 18-item multiple-choice test forms. The alternative sentence transformation method with item writers selecting the distractors proved to be somewhat superior to the objective-based test form. The study demonstrates the importance of attending to the method of item writing when potential bias is of concern in the construction of achievement tests. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ITEM BIAS DETECTION AND ITEM WRITING TECHNOLOGY

Gale H. Roid
Western Psychological Services
Los Angeles, California

Cathy L. W. Wendler
Western Psychological Services
and
University of California, Los Angeles

ED230564

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gale H. Roid

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the meeting of the American Educational Research Association, Montreal, April, 1983. Correspondence can be sent to Dr. Roid, Director of Research, WPS, 12031 Wilshire Blvd., Los Angeles, California 90025.

TM 830 220

BEST COPY AVAILABLE

BEST COPY AVAILABLE

ABSTRACT

The development of the emerging technology of item writing was motivated in part by the desire to reduce potential subjectivity and bias between different item writers who attempt to construct parallel achievement tests. The present study contrasts four test forms constructed by the combined efforts of six item writers using four methods of item writing: (1) informal and subjective, (2) objective-based, (3) sentence transformation with writer's choice of distractors, and (4) sentence transformation with algorithmic creation of distractors. Three item bias detection techniques were used to examine potential sex bias in each test form: (1) delta plots of transformed item difficulty in male and female groups, (2) Scheuneman's chi-square-correct, and (3) internal consistency estimates (alpha) for each test form in each group. The basic data of the study were from 218 4th- and 5th-grade students who were randomly assigned the four 18-item multiple-choice test forms (approximately 25-30 males or females completed each test form). The tests were designed to assess learning of basic terminology and concepts from a high-interest science article on sharks. Results of the study showed that the test form constructed by the sentence transformation method in which distractors were assembled by a clerical method produced several biased items, as determined by each of the bias detection techniques. However, the alternative sentence transformation method with item writers selecting the distractors proved to be somewhat superior to the objective-based test form, thus demonstrating the elimination of item writer bias by using an item-writing technology that allows for a controlled degree of human intervention in the item construction process. The study demonstrates the importance of attending to the method of item writing when potential bias is of concern in the construction of achievement tests.

BEST COPY AVAILABLE

Item Bias Detection and Item Writing Technology

A number of measurement theorists have suggested that the method of item writing used by a test developer has consequences for the amount of personal bias or idiosyncratic variation that is injected into an achievement test (Anderson, 1972; Berk, 1980; Bormuth, 1970). This concern has resulted in the emergence of a technology of item writing (Roid & Haladyna, 1982) in which a number of item-writing methods have been proposed that are based on specific and documentable rules by which item writers compose test items. One of the consequences of such methods is that variations in item difficulty between item writers can be brought under control (Roid & Haladyna, 1978).

The present paper examines sex bias using three approaches to bias detection: (1) differences in item difficulty for boys and girls using Angoff's (Angoff & Ford, 1973) delta-plot method, (2) group differences in correct-answer probabilities using Scheuneman's (1979) method; and (3) differences in the test's internal consistency (i.e., alpha coefficients) for boys and girls. Six different test forms were used in the study. Item distractors for the test forms were developed using contrasting methods of item writing: informal, objective-based, and two variations of sentence-transformation techniques.

Item bias, in the current study, was seen as a method of detecting flaws in the items' formats, directions, or distractors. Some researchers have used judgmental methods for establishing item bias (Tittle, 1980, 1982), while other researchers have focused on statistical techniques for detecting bias. For example, differences in item difficulty between groups have been used to indicate bias, as illustrated in Angoff's method of transformed item difficulty (Angoff & Ford, 1973; Angoff & Sharon, 1974). Others have examined the probability of a correct response for each group as a means of assessing bias, using item characteristic curve methods (Green & Draper, 1972; Lord, 1977; Wright, Mead, & Draba, 1976) or chi-square methods (Scheuneman, 1979). Finally, the relationship among items or items and total score have been examined to establish bias, using such methods as factor analysis (Gorsuch & Yagel, 1981; Green & Draper, 1972; Merz, 1978), regression analysis (Raju & Normand, 1982), and phi coefficients (Wendler & Roid, 1981).

It is expected that item bias detection methods generalize across many different types of instruments regardless of the purpose of the test or how its items were developed. While studies have examined the efficacy of different item bias detection approaches using the same test and/or set of items (Ironson & Subkoviak, 1979; Plake, 1980; Rudner, Getson & Knight, 1980a, 1980b), little has been done to examine the generalizability of item detection techniques across different test forms.

Method

Subjects

Eight classrooms of elementary school students in grades 4 and 5 in three public schools located near Salem, Oregon, were subjects for

BEST COPY AVAILABLE

the study. The schools were located in an industrial-based town of approximately 10,000 population, representing the middle range of socioeconomic status. A total of 218 students participated, 115 boys and 103 girls. Item response data were collected as part of a study of item-writing methods (Roid, Haladyna, & Shaughnessy, 1980). Table 1 shows the number of males and females assigned to each test form and the means and standard deviations of the total raw scores for each test.

Insert Table 1 about here

Procedure

Testing Procedures

Students were given 18-item multiple-choice tests before and after studying a high-interest science article on sharks, taken from a popular children's wildlife magazine. Teachers in each classroom were given packets of materials for each student, containing a pretest, a copy of the article, and a alternate-form posttest to be distributed one week later. The article had a readability level of beginning fifth grade. Instructional methods varied from a single reading of the story (e.g., by the teacher), to an integration of the story into a larger instructional unit covering one week of science lessons. Table 1 shows the means and standard deviations for pretest and posttest administrations of each test form. Also shown are the numbers of males and females given each test form on pretest and posttest. No subjects were given the same test form twice, but rather, test forms were randomly assigned for use in either pretest or posttest occasions. The purpose of including pretest data in the current study was to magnify any possible flaws in items due to guessing, independent of instruction or reading of the assigned story. Pretest and posttest data were merged for the calculation of item bias statistics in order to provide as wide a range of item performance as possible.

Item Writing Methods

Six item writers, three researchers, and three teachers trained in item writing, wrote items for the science article using four contrasting methods. This resulted in four brief test forms, each with 18 4-option multiple-choice items. Form 1 was an informal method in which each item writer constructed three items of their choice. Form 2 was based on six objectives, such as "Describe shark's food" written by the first author (three items from each writer were randomly assigned to this test form). Forms 3 and 4 were constructed using the method described by Roid and Haladyna (1982, pp. 98-108), which refined the Bormuth (1970) procedure of transforming sentences from prose material into test questions. Important nouns from key sentences in the article (as identified by an independent group of 17 elementary school teachers) were the correct answers to each question in Forms 3 and 4. However, each form differed in the method by which distractors were chosen. Form 3 used a procedure in which item writers chose their own distractors, with the restriction

BEST COPY AVAILABLE

that they be a single word. Form 4 distractors were generated by an algorithmic method (Roid & Haladyna, 1982, pp. 106-107) in which nouns of the same semantic category as the answer noun, taken from elsewhere in the article, were used as wrong-answer alternatives. Thus, Form 4 was an automated, clerical method of item writing in which sentences were transformed by specific rules and distractors chosen at random from lists of appropriate words.

Item Bias Detection Methods

Three basic techniques of item bias detection were investigated as suggested by researchers for situtations in which small samples of subjects are available for analysis (e.g., Berk, 1982). The first technique is the well-known delta-plot method described by Angoff (Angoff & Ford, 1973; Angoff, 1982). The purpose of the method is to determine the degree to which the item difficulties of each item are similar in each group. The proportion of male and female students who correctly answer each item was calculated, the proportions standardized with reference to the cumulative normal distribution (on a scale with a mean of 13 and a standard deviation of 4, with values called "deltas"). The delta values for each item were then plotted on a scatter plot with one axis representing the deltas of males and the other axis the deltas for females. Because of the small number of items in each test, it was possible to identify outliers by visual inspection. The overall dispersion of points from the best-fitting regression line was summarized by the correlation between the deltas of the two groups. Although the two groups may differ somewhat in level of ability, the points should fall in an ellipsoidal area bisected by the regression line, unless certain items change qualitatively or in their rank order between the two groups.

Although it is recognized that the delta plot method has certain limitations, such as sensitivity to differences in ability levels between groups or differences in item discrimination (Angoff, 1982), it remains a widely used and well-understood method. When used in combination with other methods, such as the two additional techniques described below, it seemed to be a reasonable alternative for the present study. Sample sizes were not judged to be sufficient in size to allow for matching of samples on ability or for correcting the deltas through division by a standardized discrimination index as suggested by some researchers (Angoff, 1982).

The second technique used was the chi-square method proposed by Scheuneman (1979) and recommended for small samples by Ironson (1982). The logic of this approach is that an item is considered unbiased if individuals with equal ability from different groups have the same probability of answering the item correctly. The chi-square-correct statistic can be viewed as a rough approximation to the more sophisticated latent-trait (item characteristic curve, ICC) approaches (Ironson, 1982) which examine the shape of the ICC curves for each group. The ICC methods such as the Rasch or 3-parameter models could not be appropriately applied to the present study because of the relatively small sample sizes available for each test form. The chi-square index was computed for this study by dividing the total score distributions for each test form into three seg-

BEST COPY AVAILABLE

ments, represent low, average, and high ability students. Then, the proportions of males and females responding correctly to each item were calculated for each ability group.

It should be noted that several investigators have shown that the Scheuneman chi-square-correct index in not necessarily distributed as the chi-square statistic because only the correct responses are considered. However, the index has been shown to be highly correlated with more complex ICC indexes in several empirical comparisons (e.g., Rudner, Getson, & Knight, 1980b).

The final technique, comparison of the internal consistency of test forms, was used as a way of summarizing the cumulative effects of potential error due to differences in the covariance between items within the male or female subsamples. The approach taken in the current paper was to compare coefficient alpha (KR-20) for each test form as recommended by Reynolds (1982). Reynolds suggests that alpha provides information about the degree of error due to domain sampling that may be detectable by examining the differences between alpha coefficients computed separately on the two subsamples. The method of Feldt (1969) was used to calculate F-tests for the differences between alphas.

Results

Results for each of the item bias detection methods will be discussed in turn. First, the delta plots for each of the four test forms were visually inspected and the correlations between the deltas in each group calculated. As shown in Table 2, Form 4, the sentence-transformation, algorithmic item writing method, showed the greatest dispersion in its delta plot, with a correlation of only .544. Six widely dispersed items were identified, three of which were easier for boys and three easier for girls. Inspection of these 6 potentially biased items in Form 4 indicated that two of the items were flawed by having multiple correct answers, and it appeared that boys were attracted to plausible incorrect answers from background information about animals obtained independently of the story used in the experiment. Girls appeared to have attended more to details of the story and were not as distracted by commonly plausible incorrect answers.

Insert Table 2 about here

The delta plots for Forms 1, 2 and 3 provide some interesting contrasts, although the differences are subtle. The informal item writing method of Form 1 did surprisingly well, and showed the greatest range of item difficulties. The test form with the least dispersion was Form 3 (r=.881), the sentence-transformation, writer's choice method which had a more narrow range of item difficulties (mostly easier items compared to Form 1 or 2), except for a single item which was only answered correctly by 1 male student and by none of the females. Inspection of this difficult item in Form 3 revealed that it was an oddly worded item with very lengthy distractors measuring a subtle fact about basking sharks. Form 3, the objective-

BEST COPY AVAILABLE

based item writing method, produced a relatively more dispersed delta plot than Forms 1 or 3. Inspection of the four most dispersed items revealed that girls found certain items easier than boys, particularly those dealing with the feeding habits or birth of newborn sharks, implying a possible increased attention by females to these details in the reading materials.

The results of Scheuneman's chi-square analysis are presented in Table 3. If one can assume that the index is distributed as chi-square approximately, three items would be found to have statistically significant chi-square values in Form 4, one biased item in Form 2, and no items in Forms 1 or 3. This procedure may be a conservative one which detects only highly biased items. Inspection of the individual items revealed that guessing factors appeared to play a greater role for girls than for boys. Girls who obtained lower total scores were able to guess the correct answer more frequently than boys for the items detected as biased. All three of the biased items in Form 4 had plausible incorrect answers that boys may have checked due to background information about fish or hunting, independent of the details of the story, whereas the girls appeared to choose alternatives that were more plausible in the context of the story.

Insert Table 3 about here

Table 4 presents the results of the comparisons for the internal consistency estimates for each test form. F-tests on the differences between subsample alphas indicated a significant difference on only one test form, Form 4, the sentence-transformation, algorithmic method. The alpha for the female subsample on this test form was only .414, the lowest of all the alphas computed. Inspection of the item-total-score correlations and the item intercorrelations revealed two items with negative correlations with total score, and relatively high negative intercorrelations with other items. These two items had also been identified by the delta plot and the chi-square index as biased items.

Insert Table 4 about here

Discussion

Each of the three item bias detection techniques used in the study were consistent in identifying the most biased (Form 4) and the least biased (Form 3) of the test forms constructed by alternative item writing methods. Form 3, the sentence-transformation method with writers choosing the one-word distractors, had the highest correlation between delta values, the highest alpha values with no differences between alphas computed on male and female subsamples, and no items identified as significantly biased by the chi-square technique. Form 4, also a sentence-transformation method, but with distractors created clerically from lists of words taken from the prose material in the experiment, was shown to be inferior on all criteria.

Although the experiment represents the contrast of item writing

BEST COPY AVAILABLE

methods used in one content area (science/zoology), at one educational level (grades 4-5), some tentative conclusions can be cautiously suggested. It appears that a moderate amount of control over the item writer's construction of items, via specific rules for selection of key elements (based on teacher's selection of key words in a prose passage), while still allowing the item writers to choose their own distractors, proved to be slightly superior to an informal technique in the present experiment. One might argue, alternatively, that the informal method of Form 1, which is the conventional way in which instructors write test items which are subjectively felt to be important, was the most cost-effective method because it was nearly as unbiased and had nearly the same degree of internal consistency as the more complex sentence-transformation technique. However, another interpretation is that the rule-based sentence-transformation method may be implemented in a computer-aided item-writing system that could ultimately allow for the creation of hundreds of items in a semi-automated fashion. If such a system were possible, the sentence-transformation method with writer's intervention in the choice of distractors may prove to be the most cost-effective.

Also, the item writing method of Form 3 appeared to control item bias more carefully than an objective-based method. The experiment may suggest that selection of the objectives themselves may be the point at which bias can enter the test development process, as evidenced by the subjective judgement that certain items in Form 2, measuring the objectives of feeding and birth of sharks, were more easily answered by females who may have somewhat more attentiveness to this content area.

The results of the study reveal the importance of examining item writing methods as a factor in determining the quality of achievement tests and the items that compose them. It is apparently not enough to simply review items, or to assume that specification of objectives to guide item writing will control for potential item bias. Rather, all items, and perhaps items grouped by instructional objective, need to be subjected to the recommended item bias detection methods prior to identifying potentially flawed items. Concern for the problem of distorting a domain-referenced item pool by the deselection of items can be alleviated by treating the item bias review process as a means for editing and improving items, rather than a means for eliminating troublesome items. Also, as suggested by the present study, the method of item writing and the particular rules employed to guide item writers can be investigated through comparative studies so as to modify any methods of item construction that may be producing an unacceptable number of biased items for a particular content area or type of subject.

BEST COPY AVAILABLE

References

Anderson, R.C. How to contstruct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Angoff, W.H. Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore: The Johns Hopkins University Press, 1982.

Angoff, W.H., & Ford, S.T. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

Angoff, W.H., & Sharon, A.T. The evaluation of differences in test performance on two or more groups. Educational and Psychological Measurement, 1974, 34, 807-816.

Berk, R.A. (Ed.). Criterion-referenced measurement. Baltimore: The Johns Hopkins University Press, 1980.

Berk, R.A. (Ed.). Handbook of methods for detecting test bias. Baltimore: The Johns Hopkins University Press, 1982.

Bormuth, J.R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Feldt, L.S. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 1969, 34, 363-373.

Gorsuch, R.L., & Vagel, J.C. Item factor analysis. Paper presented at the meeting of the American Educational Research Association, Los Angeles, April 1981.

Green, D.R., & Draper, J.F. Exploratory studies of bias in achievement tests. Monterey, Ca.: McGraw-Hill, 1972.

Ironson, G.J. Use of chi-square and latent trait approaches for detecting item bias. In R.A. Berk (Ed.), Handbook of methods for detecting item bias. Baltimore: The Johns Hopkins University Press, 1982.

Ironson, G.J., & Subkoviak, M.J. A comparison of several methods of assessing item bias. Journal of Educational Measurement, 1979, 16, 209-255.

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Merz, W.R. Estimating bias in test items utilizing principal components analysis and the general linear solution. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976.

Plake, B.S. A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. Educational and Psychological Measurement, 1980, 40, 397-404.

Raju, N.S., & Normand, J. The regression bias method: A unified approach for detecting item bias and selection bias. Paper presented at the meeting of the National Council on Measurement in Education, New York, March 1982.

Reynolds, C.R. Methods for detecting construct and predictive bias. In R.A. Berk (Ed.), Handbook of methods for detecting item bias. Baltimore: The Johns Hopkins University Press, 1982.

Reid, G.R., & Haladyna, T.M. A comparison of objective-based and mod-

BEST COPY AVAILABLE

ified-Bormuth item writing techniques. Educational and Psychological Measure, 1978, 35, 19-28.

Roid, G.R., & Haladyna, T.M. A technology for test item writing. New York: Academic Press, 1982.

Roid, G.R., Haladyna, T.M., & Shaughnessy, J. A comparison of item-writing methods for criterion-referenced testing. Paper presented at the meeting of the National Council on Measurement in Education, Boston, April 1980.

Rudner, L.M., Getson, P.R., & Knight, D.L. Biased item detection techniques. Journal of Educational Statistics, 1980a, 5, 213-233.

Rudner, L.M., Getson, P.R., & Knight, D.L. A monte carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 1980b, 17, 1-10.

Scheuneman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Tittle, C.K. Judgmental methods in test development. Paper presented at the Johns Hopkins University National Symposium on Educational Research, Washington, D.C., November 1980.

Tittle, C.K. Uses of judgmental methods in item bias studies. In R.A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore: The Johns Hopkins University Press, 1982.

Wendler, C.L.W., & Roid, G.H. Reducing sex bias in a vocational interest inventory. Paper presented at the meeting of the American Psychological Association, Los Angeles, August 1981.

Wright, B.D., Mead, R., & Draba, R. Detecting and correcting test item bias with a logistic response model. (Research memorandum number 22). Chicago: Department of Education, University of Chicago, 1976.

BEST COPY AVAILABLE

Table 1

Number of Subjects, Raw Score Means, and Standard Deviations for Each Test Form

Test Form	Sex	Pretest n	Pretest Mean	Pretest SD	Posttest n	Posttest Mean	Posttest SD	Total n	Total Mean	Total SD
Form 1: Informal Method	M	14	5.43	2.28	11	10.09	4.89	25	7.48	4.28
	F	12	5.75	3.11	12	8.67	3.82	24	7.21	3.72
Form 2: Objective-Based Method	M	15	6.53	2.97	12	8.67	4.96	27	7.48	4.04
	F	12	7.67	3.23	16	10.25	3.02	28	9.14	3.32
Form 3: Sentence Transform, Writer's Choice Method	M	15	6.87	3.31	16	11.31	3.86	31	9.16	4.20
	F	12	7.08	4.06	14	9.93	3.81	26	8.62	4.11
Form 4: Sentence Transform, Algorithmic Method	M	18	9.72	3.82	14	10.93	4.81	32	10.25	4.25
	F	10	11.30	2.00	15	12.27	2.71	25	11.88	2.45

Note. All test forms were 18-item, multiple-choice tests with four answer alternatives.

Table 2

Interitem Correlations from Delta Plots

Test Form		Correlations
Form 1:	Informal Method	.855
Form 2:	Objective-Based Method	.757
Form 3:	Sentence Transform, Writer's Choice Method	.881
Form 4:	Sentence Transform, Algorithmic Method	.544

Table 3

Scheuneman's Chi-Square for Items in Each Test Form

Item No.	Form 1	Form 2	Form 3	Form 4
1	3.12	.37	3.65	.21
2	1.71	.23	.58	.73
3	2.17	.07	2.31	1.92
4	.60	.20	1.17	.15
5	1.98	1.19	.61	.79
6	.45	2.18	2.56	1.87
7	..77	.55	.90	.12
8	1.20	.58	.37	1.13
9	.97	.89	.85	1.35
10	.22	1.54	1.20	.26
11	1.01	.19	2.04	1.00
12	1.02	2.16	1.60	4.93*
13	2.75	4.27	1.12	.15
14	2.12	.76	2.57	4.82*
15	3.04	.96	.35	5.07*
16	2.27	3.77	2.97	2.13
17	.62	5.01*	3.06	1.09
18	.67	.19	.63	3.45
$\bar{X}$ Chi-square	1.48	1.42	1.59	1.73
SD Chi-square	.93	1.53	1.04	1.71

*$p \leq .10, \chi^2 = 4.605$

Table 4

Internal Consistency Estimates (Alphas)
and F-tests for Each Test Form

Test Form		Alphas Males	Alphas Females	F-test	df
Form 1:	Informal Method	.813	.753	1.32	23,24
Form 2:	Objective-Based Method	.795	.660	1.66	27,26
Form 3:	Sentence Transform, Writer's Choice Method	.824	.804	1.12	25,30
Form 4:	Sentence Transform, Algorithmic Method	.815	.414	3.17*	24,31

*$p \leq .01$